Shaswat Patel

Email : shaswat178@gmail.com
Mobile : +1(201) 918-9767

Education

- **New York University** — New York, NY
  Master of Science in Computer Science; GPA: 3.72/4.0 — *Sept 2024 – May 2026*
- **Netaji Subhas Institute of Technology** — New Delhi, India
  Bachelor of Engineering in Computer Engineering; GPA: 8.2/10.0 — *Jun 2018 – May 2022*

Publications

- **A novel momentum-based deep learning techniques for medical image classification and segmentation.** MICCAI MLMI Workshop
- **A Dataset for Detecting Humor in Telugu Social Media Text** ACL Workshop
- **Bias Amplification in Intersectional Subpopulations for Clinical Phenotyping by Large Language Models** Medrxiv
- **ShockModes: A Multimodal Model for Prognosticating Intensive Care Outcomes from Physician Notes and Vitals** Medrxiv
- **Rumour detection on benchmark twitter datasets using graph neural networks with data augmentation** Springer Nature Social Network Analysis and Mining Journal

Experience

- **Studio Management** — NY, USA
  Machine Learning Engineer Intern — *Jun 2025 - Aug 2025*
  - **Event Recommendation GPT**: A search and discovery utility tool for real-life events. The system consists of:
    1. **RAG:** Fine-tuned a distilled BERT model with GPT-generated metadata to support metadata-based filtering, **achieving 90% accuracy while reducing latency by 75%**. Implemented a dense retrieval pipeline leveraging OpenAI embeddings, followed by re-ranking via Jina AI's cross-encoder, incorporating multi-dimensional signals such as venue information and out-of-distribution scoring to prioritize unique and diverse events.
    2. **Memory Module: Implemented a custom memory architecture to store long-term user preferences** (e.g., preferred genres, locations). The implementation includes tool-calling for CRUD operations on memory. The current module is 85% accuracy at extracting user preferences.
    3. **Multi-Turn Summarizer:** Implemented standard techniques to support multi-turn conversation by maintaining state of the conversation via summarizer.
    4. **ReAct Tool-Calling Framework:** To support multiple queries, the **RAG pipeline is encapsulated via a ReACT tool-calling framework.** Queries include: weekly calendar creation and daily planner which includes restaurants, activities and events.
    5. **Test dataset:** Automatic pipeline to curate and store events for testing different components of RAG pipeline. qThe pipeline also includes E2E testing via GEval and DeepEvals.
- **New York University** — NY, USA
  Research Assistant(RA) — *Apr 2025 - Present*
  - **Long-horizon coding tasks (Advisor: Prof. He He)**: Current benchmarks such as SWE-Bench Verified measure only one-step resolve rates, while real-world agents must frequently modify code they previously wrote. I am developing a new benchmark that captures these long-horizon, self-referential coding scenarios and building a maintainability evaluation toolkit that assesses coding agent performance beyond simple task resolution.
  - **Multilingual Retrieval Heads (Advisor: Prof. Eunsol Choi)**: Retrieval heads are key components that help extract relevant information from long in-context texts to respond to queries. This project extends prior work into multilingual settings, studying how retrieval behavior varies based on the language of the query and context.
- **Walmart** — Chennai, India
  Software Engineer — *Jul 2022 - Jul 2024*
  - **Confluence based Chatbot**: Spearheaded the development of Confluence-integrated chatbot using Retrieval Augmented Generation(RAG) and Large Language Models(LLMs) to generate Walmart's system specific text responses. Responsible for building the backend using FastAPI(python), LangChain, and OpenAI's chatGPT API.

- **The Minion Project**: Played a key role in developing APIs for the Automated Continuous Monitoring System, affectionately known as Minion. **This effort enhanced *Time to Detect*(TTD) by 50%**, eventually leading to its adoption as a framework across all critical Walmart systems. The APIs are developed using FastAPI(python).
- **Project Galaxy**: Spearheaded the development of the frontend of the Project Galaxy which focused on quick recovery from a bug across multiple systems(and teams). The project involved more than five dependent different teams. **This effort enhanced *Time to Recover*(TTR) by 30%.**
- **Bravo award**: Honored with the Bravo Award for outstanding contributions.

- **Tavlab** Delhi, India
  Research Associate (RA) *Nov 2020 - Oct 2023*
  - **Addressed biomedical challenges**: Applied NLP (BioBERT, LLaMA, Mistral) embeddings and Machine Learning (Logistic Regression, SVM, Gradient Boosting) to solve biomedical challenges. Focused on tasks including sepsis prediction, abnormal shock index prediction, mortality estimation, and length of stay analysis. Implemented modeling sharding on multi-GPU setup and increased inference speed by 15%. Acheived state-of-the-art results.
  - **COVID Gene Sequencing**: Build a transformer based model to understand genomic sequences for covid strains. Using the next-sentence prediction task, transformer the DNA sequence into codons and build a vocabulary of all the possible codons. Then trained the model on a large corpus of existing known DNA sequences. Build a regression model on top of this model to predict the number of patients in the next 1-, 2-, and 3- months.
- **MIDAS LAB** Delhi, India
  Research Associate (RA) *Nov 2020 - Nov 2022*
  - **video2vec-U**: Built an end-to-end, unsupervised Visual Speech Recognition system that combines computer vision and natural language processing models, inspired by the wav2vec-U framework. Finetuned visual-encoder(CNN and ViT-based model) and Language Model(n-gram, LSTM, and transformer) from scratch to capture visual features(mainly facial features) of the speaker. Used multi-GPU setup for finetuning on large-scale video dataset.
- **Noble Missions for Change Initiative (United Nation)** Remote
  Web developer *Feb 2020 - Apr 2020*
  - **Frontend development**: Developed the frontend of a web application connecting remote Nigerian schools with teachers globally. This was done as part of United Nation outreach program.
  - **Project Management**: Coordinated with stakeholders and cross-functional teams to deliver the project.

Projects

- **Attention-Aware DPO**: Large Language Vision Models(LLVMs) suffer from inadequacies when multi-image query is prompted. In thie project, we modified the existing DPO loss(used for human preference alignment) to include a grounding loss based on model's attention mechanism.
- **Multilingual Retrieval Heads**: Large Language Models(LLMs) are effective at extracting key information from a long in-context prompt. But how did LLM excel at this task? We mechanistically explain this phenomenon in multilingual setting. Currently in work with Professor Eunsol Choi.
- **GPU Accelerated Limit Order Book Heads**: Implemented a high-performance Limit Order Book (LOB) engine on GPUs using C++ CUDA, demonstrating that traditionally CPU-bound LOB operations can be efficiently parallelized. Achieved **120x speedup gains over CPU baselines** and extended the system with APIs enabling large-scale multi-agent RL training. Developed as part of the NYU GPU Computing course.

Teaching

- **CSCI-UA.0480-052: Algorithmic Problem Solving** NYU
  Teaching Assistant *Sept 2024 - Dec 2024*
- **CSCI-UA.0102-001: Data Structures** NYU
  Teaching Assistant *Jan 2025 - May 2025*
- **CSCI-UA.0480-052: Algorithmic Problem Solving** NYU
  Course Assistant *Sept 2025 - Dec 2025*
- **CSCI-UA.0102-001: Data Structures** NYU
  Teaching Assistant *Sept 2025 - Dec 2025*

Volunteering

- **Lions Club International** India
  Volunteer *May 2020 - Jun 2020*
- **V care foundation** India
  Volunteer *Jun 2020 - Jul 2020*